BABU BANARASI DAS ENGINEERING COLLEGE

(AKTU Code: 508)

Affiliated to Dr. A.P.J. Abdul Kalam Technical University, Lucknow

DEPARTMENT OF INFORMATION TECHNOLOGY

B.TECH.(IT4) FINAL YEAR

ACADEMIC SESSION 2023-24 (ODD SEMESTER)

MINI PROJECT SYNOPSIS

ON

LOAN PREDICTION SYSTEM

Submitted By	**Submitted To**
Alka Gupta	Mrs. Priyanka Gupta
Roll No. 2005080130007	(Assistant Professor)

ACKNOWLEDGMENT

I am pleased to acknowledge my sincere thanks to the Head of Department of Information Technology, Dr. Santosh Shukla, and the assistant professor Mrs. Priyanka Gupta for their kind encouragement in doing this project and for completing it successfully.

I would like to express my sincere and deep sense of gratitude to Internshala for consistently guiding me during the entire training session and helping me to understand and complete the project on the Loan Prediction System using Machine Learning.

I am grateful to them for providing me necessary support and details at the right time during the progressive reviews.

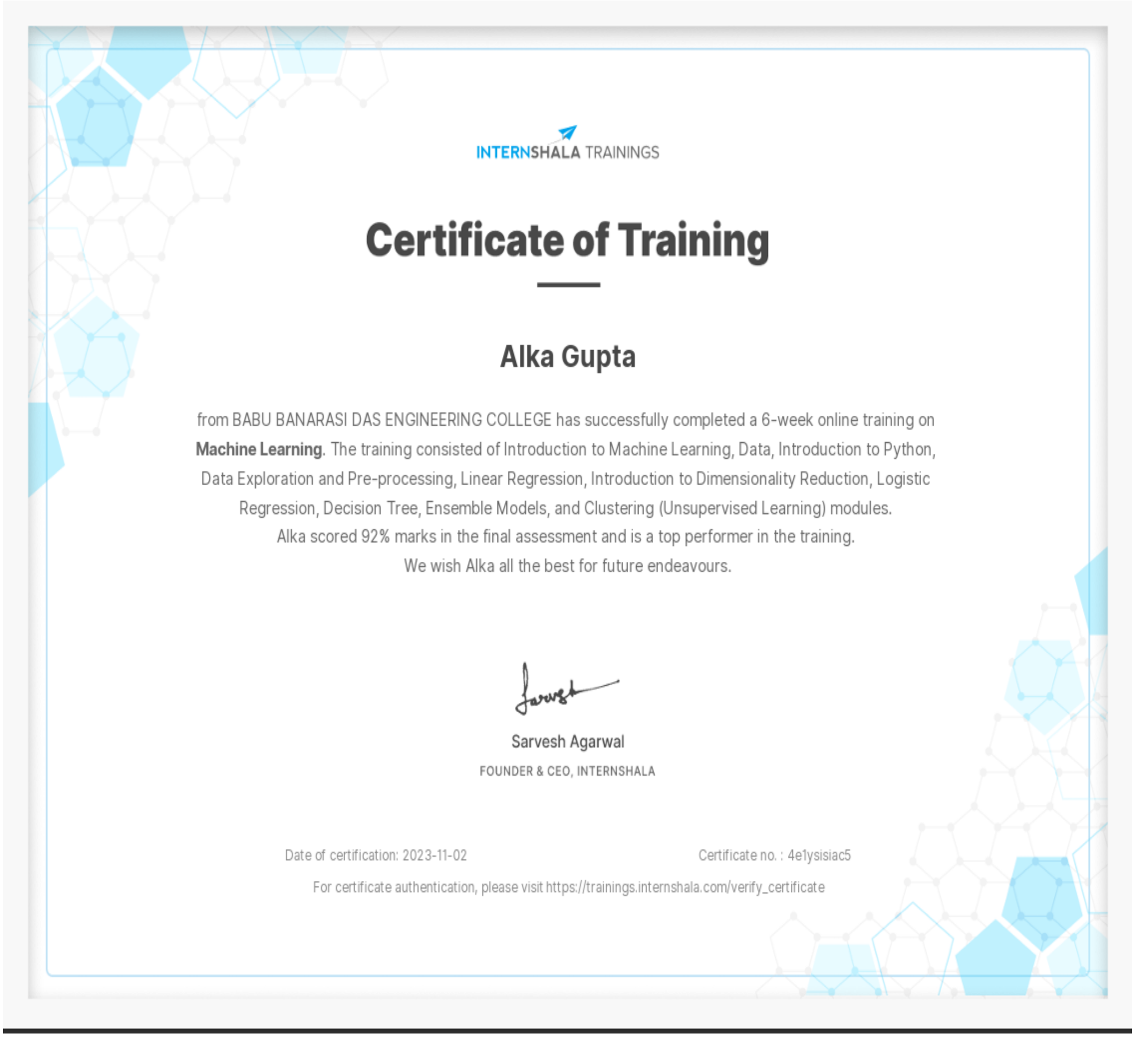
I wish to express my thanks to all teaching and non-teaching staff members of the Department who were helpful in many ways for the completion of the project.

Last but not least I would like to thank my friends who helped in making this project.

ABSTRACT

Technology has boosted the existence of humankind and the quality of life they live. Every day we are planning to create something new and different. We have a solution for almost every problem, we have machines to support our lives and make us somewhat complete in the banking sector as well. Banks are making a major part of their profits through loans. Though a lot of people are applying for loans, it's hard to select the genuine applicant, who will repay the loan. While doing the process manually, a lot of misconceptions may happen to select a genuine applicant. Therefore, here developing a loan prediction system using machine learning, so the system automatically selects the eligible candidates. This is helpful to both bank staff and applicants. The time for the sanction of a loan will be drastically reduced. In this report, we are predicting the loan data by using some machine learning algorithms.

CERTIFICATE OF COMPLETION



INTERNSHALA TRAININGS

Certificate of Training

Alka Gupta

from BABU BANARASI DAS ENGINEERING COLLEGE has successfully completed a 6-week online training on **Machine Learning**. The training consisted of Introduction to Machine Learning, Data, Introduction to Python, Data Exploration and Pre-processing, Linear Regression, Introduction to Dimensionality Reduction, Logistic Regression, Decision Tree, Ensemble Models, and Clustering (Unsupervised Learning) modules.
Alka scored 92% marks in the final assessment and is a top performer in the training.
We wish Alka all the best for future endeavours.

Sarvesh Agarwal
FOUNDER & CEO, INTERNSHALA

Date of certification: 2023-11-02

Certificate no. : 4e1ysisiac5

For certificate authentication, please visit https://trainings.internshala.com/verify_certificate

INDEX

INTRODUCTION

As the data is increasing daily due to digitalization in the banking sector, people want to apply for loans through the Internet. Artificial intelligence (AI), as a typical method for information investigation, has gotten more consideration increasingly. Banks are facing a significant problem in the approval of the loan. Daily there are so many applications that are challenging to manage by the bank employees, and also the chances of some mistakes are high.

Most banks earn profit from the loan, but it is risky to choose deserving customers from the number of applications. One mistake can make a massive loss to a bank. Loan distribution is the primary business of almost every bank. The main portion of the bank's profit directly comes from the profit earned from the loans.

Though the bank approves a loan after a regress process of verification and testimonial still there's no surety whether the chosen hopeful is the right hopeful or not.
The bank authorities complete all other customer's other formalities on time, which positively impacts the customers. The best part is that it is efficient for both banks and applicants.

PROPOSED SYSTEM

To deal with the problem, we developed automatic loan prediction using machine learning techniques. We will train the machine with the previous dataset so that the machine can analyze and understand the process. Then the machine will check for eligible applicants and give us results.

Features of the Proposed System

- The time for loan sanctioning will be reduced.
- The whole process will be automated, so human error will be avoided.
- Eligible applicants will be sanctioned loan without any delay.

Hardware and Software Specification

Hardware Specification

1. Operating System: Independent (MS Windows, macOS, Linux)
2. Memory: 128 GB
3. RAM: 4 GB
4. Internet Connectivity: Required

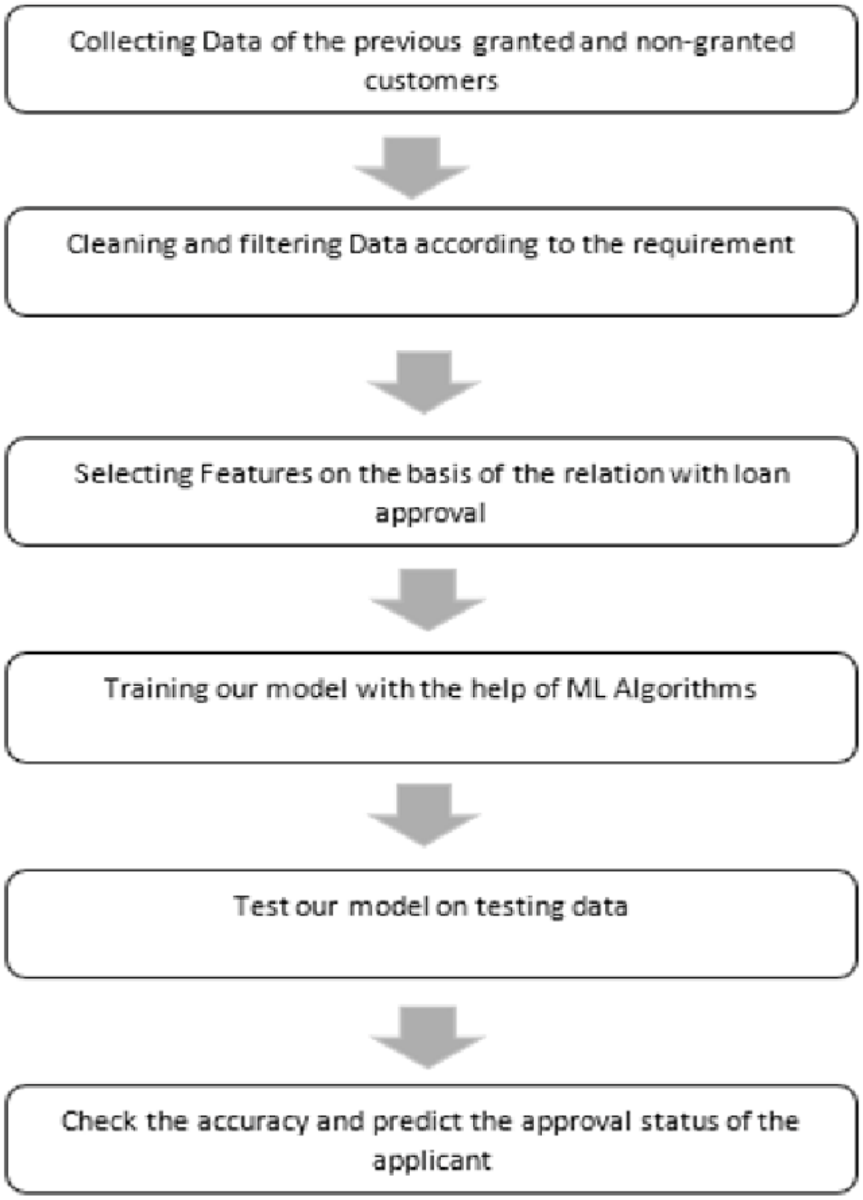
Software Specification

1. Coding Language: Python
2. IDE: Jupyter Notebook, VS Code

Library Used

1. Pandas
2. Numpy
3. Matplotlib
4. Sklearn

BLOCK DIAGRAM


Collecting Data of the previous granted and non-granted customers
Cleaning and filtering Data according to the requirement
Selecting Features on the basis of the relation with loan approval
Training our model with the help of ML Algorithms
Test our model on testing data
Check the accuracy and predict the approval status of the applicant

PROPOSED METHODOLOGY

1. **Collect the data:** First, the data is collected.

2. **Prepare the input data**: This step is done by the original owners of the dataset, and the composition of the dataset.

3. **Analyze the input data**: Understand the relationship among different features. A plot of the core features and the entire dataset. The dataset is further split into 2/3 for training and 1/3 for testing the algorithms. Furthermore, to obtain a representative sample, each class in the full dataset is represented in about the right proportion in both the training and testing datasets.

4. **Train the algorithm:** The various classification algorithms are trained using a different set of data.

5. **Test the algorithm:** The various algorithms are used to predict the effectiveness of the algorithm on the test dataset. In evaluating the performance of the classification algorithms, it includes accuracy, precision, recall, and many more.

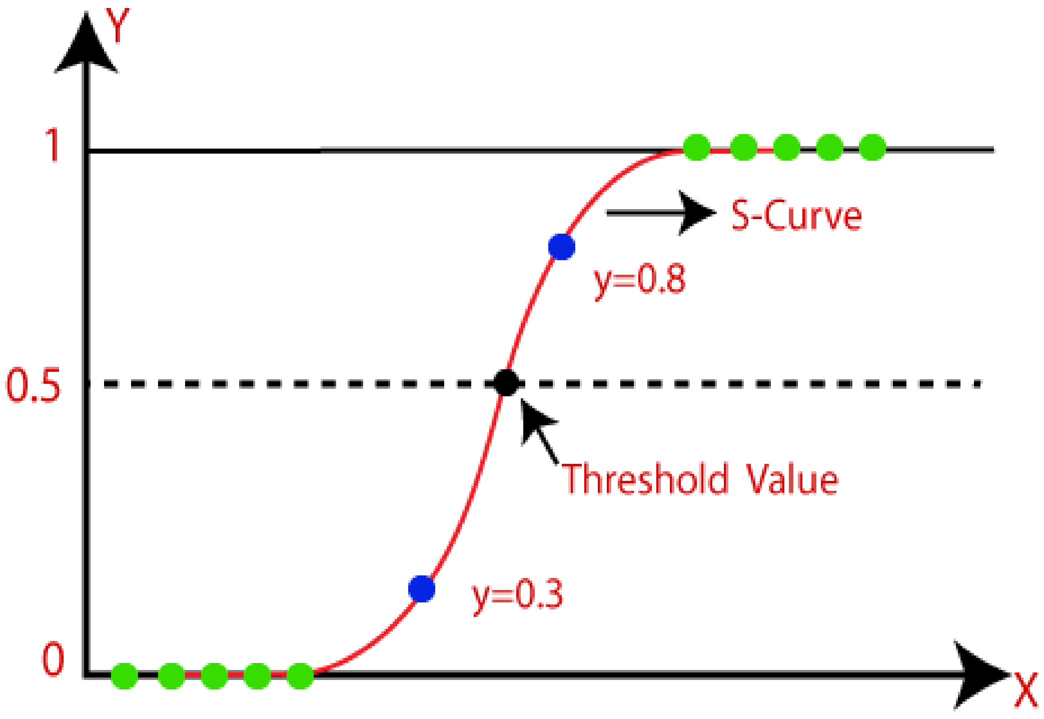
6. These values are calculated using the Python scikit learn tool with input values as the entities of the confusion matrix. A ‘positive’ instance refers to no (signifying there will not be a default in the payment of the loan) whereas the ‘negative’ instance refers to yes (signifying there will be a default in the payment of the loan).

ALGORITHM

A. Logistic regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- Logistic Regression is much similar to Linear Regression except for how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it can provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

The below image shows the logistic function:



$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- We know the equation of the straight line can be written as:

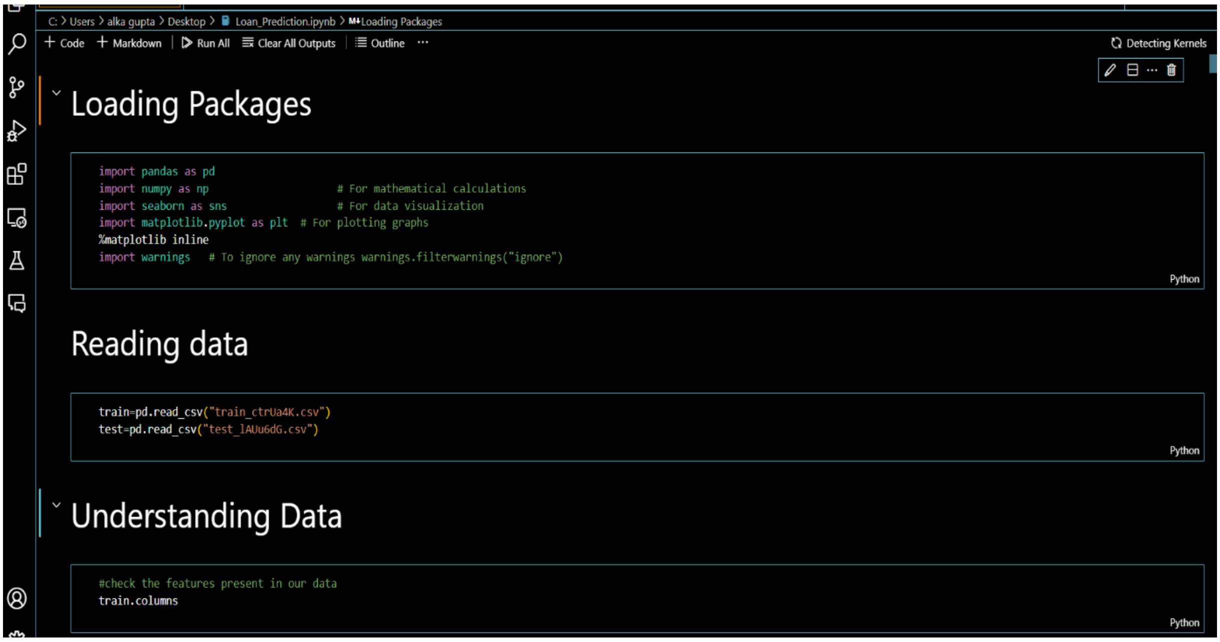
- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$\frac{y}{1-y}$; **0 for y= 0, and infinity for y=1**

PROPOSED WORK

- Import all the required Python modules.
- Import the database for both TESTING and TRAINING.
- Check if any NULLVALUES exist.
- If NULLVALUES exits, fill the table with the corresponding coding.
- Exploratory Data Analysis for all ATTRIBUTES from the table
- Plot all graphs using MATPLOTLIB module.
- Send that output to CSV FILE.

SCREENSHOTS



C: > Users > alka gupta > Desktop > Loan_Prediction.ipynb > M↓Loading Packages
+ Code + Markdown | ▷ Run All ☰ Clear All Outputs | ☰ Outline ···
Detecting Kernels

Loading Packages

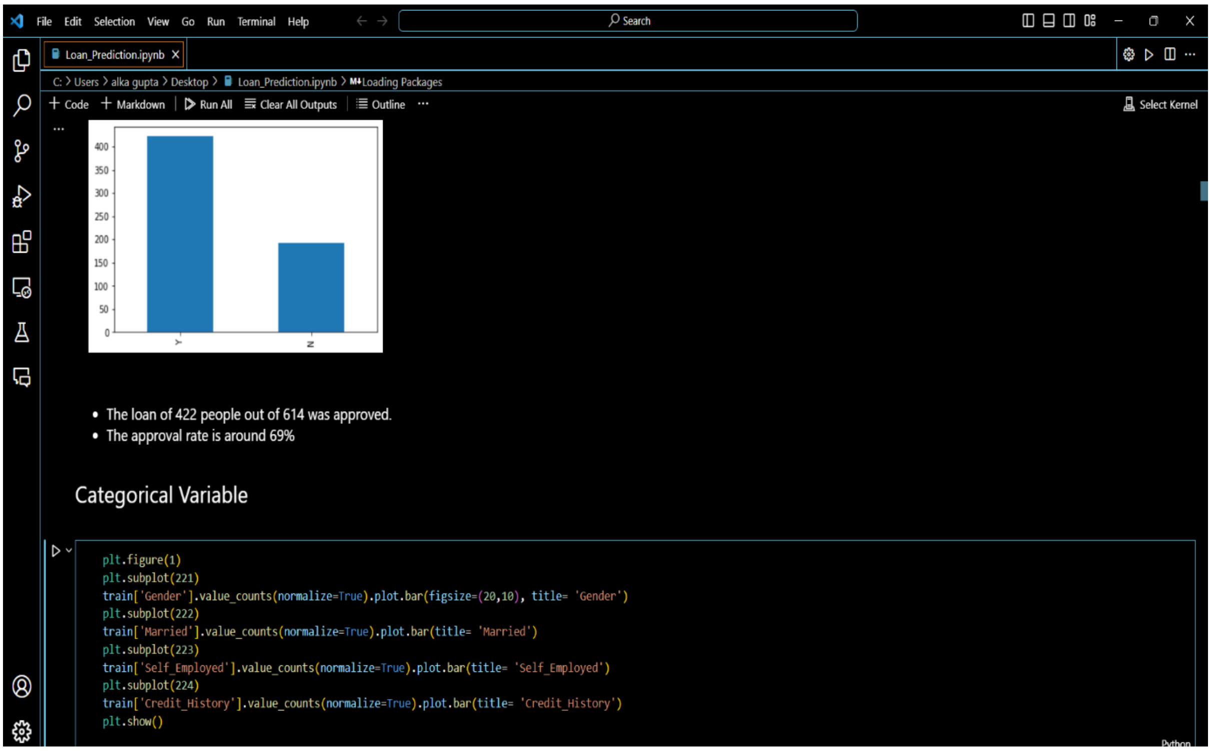
```
import pandas as pd
import numpy as np                    # For mathematical calculations
import seaborn as sns                 # For data visualization
import matplotlib.pyplot as plt  # For plotting graphs
%matplotlib inline
import warnings   # To ignore any warnings warnings.filterwarnings("ignore")
```

Reading data

```
train=pd.read_csv("train_ctrUa4K.csv")
test=pd.read_csv("test_lAUu6dG.csv")
```

Understanding Data

```
#check the features present in our data
train.columns
```

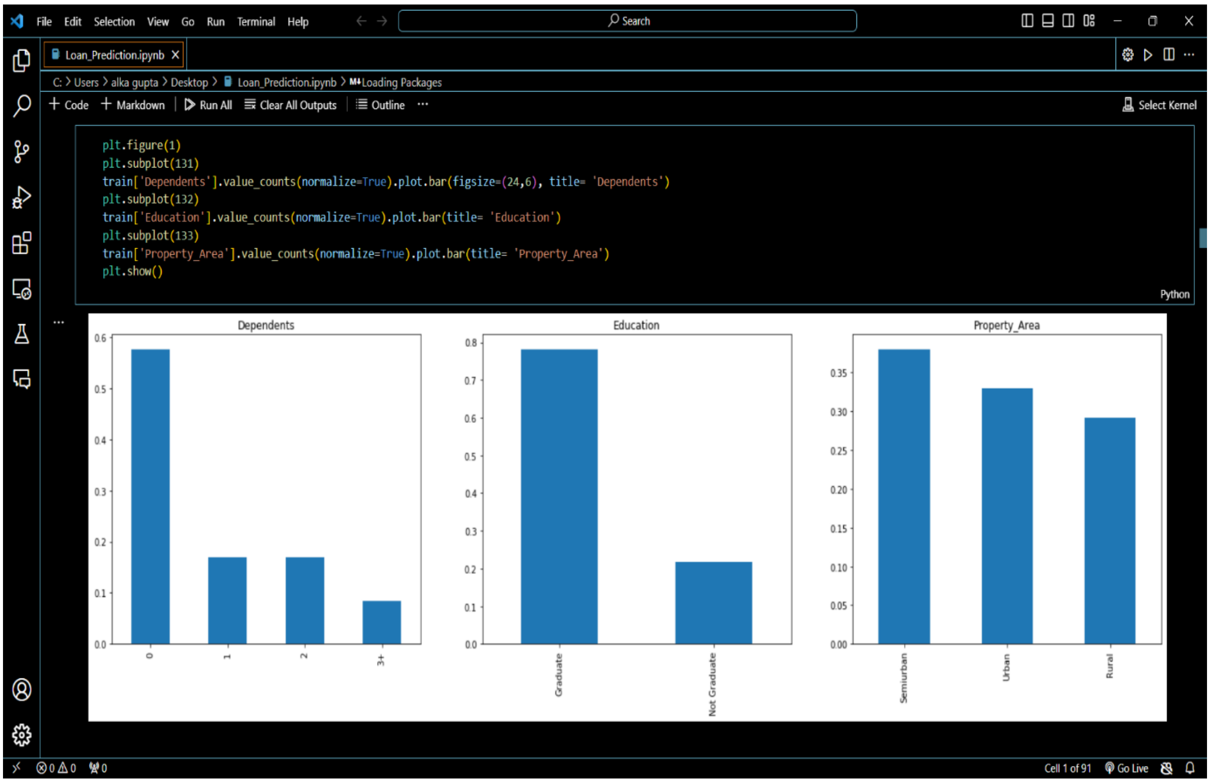


File Edit Selection View Go Run Terminal Help
Search
Loan_Prediction.ipynb
C: > Users > alka gupta > Desktop > Loan_Prediction.ipynb > M↓Loading Packages
+ Code + Markdown | ▷ Run All ☰ Clear All Outputs | ☰ Outline ···
Select Kernel
400
350
300
250
200
150
100
50
0
Y
N

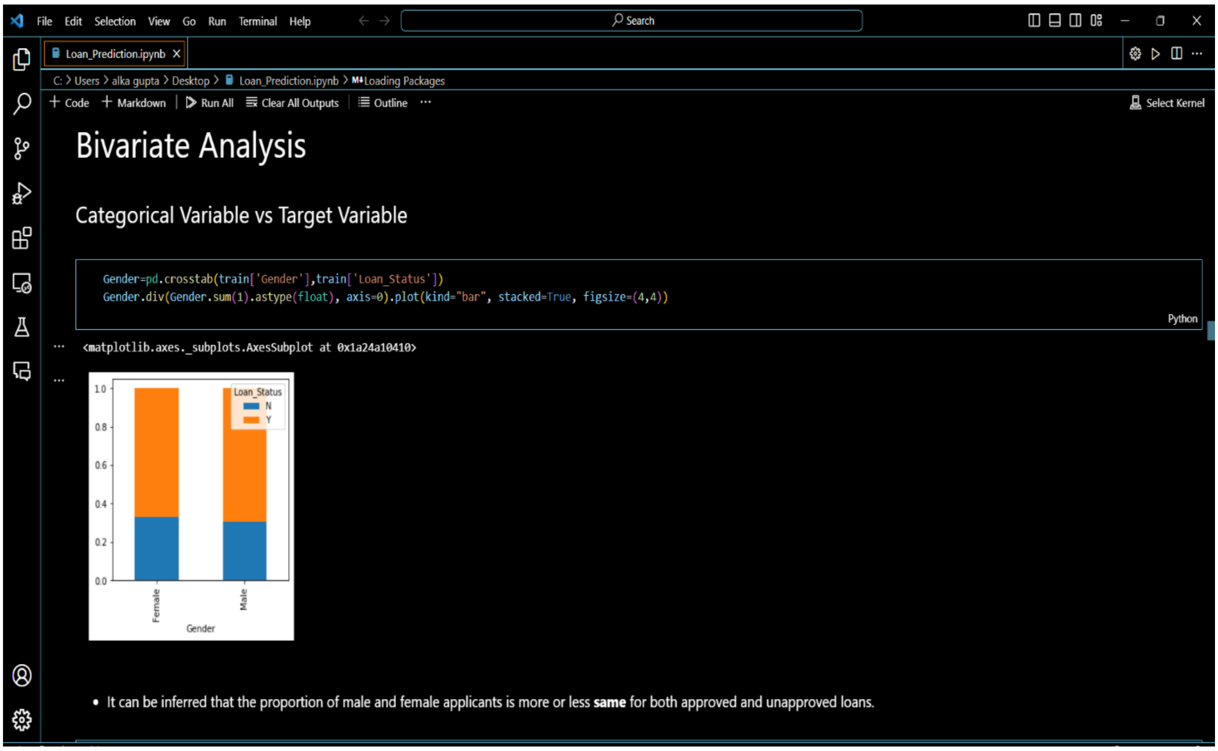
- The loan of 422 people out of 614 was approved.
- The approval rate is around 69%

Categorical Variable

```
plt.figure(1)
plt.subplot(221)
train['Gender'].value_counts(normalize=True).plot.bar(figsize=(20,10), title= 'Gender')
plt.subplot(222)
train['Married'].value_counts(normalize=True).plot.bar(title= 'Married')
plt.subplot(223)
train['Self_Employed'].value_counts(normalize=True).plot.bar(title= 'Self_Employed')
plt.subplot(224)
train['Credit_History'].value_counts(normalize=True).plot.bar(title= 'Credit_History')
plt.show()
```



```
plt.figure(1)
plt.subplot(131)
train['Dependents'].value_counts(normalize=True).plot.bar(figsize=(24,6), title= 'Dependents')
plt.subplot(132)
train['Education'].value_counts(normalize=True).plot.bar(title= 'Education')
plt.subplot(133)
train['Property_Area'].value_counts(normalize=True).plot.bar(title= 'Property_Area')
plt.show()
```


File Edit Selection View Go Run Terminal Help
Search
Loan_Prediction.ipynb
C: > Users > alka gupta > Desktop > Loan_Prediction.ipynb > M↓Loading Packages
+ Code + Markdown | Run All Clear All Outputs | Outline ...
Select Kernel
Python
1.0
0.8
0.6
0.4
0.2
0.0
Loan_Status
N
Y
Female
Male
Gender

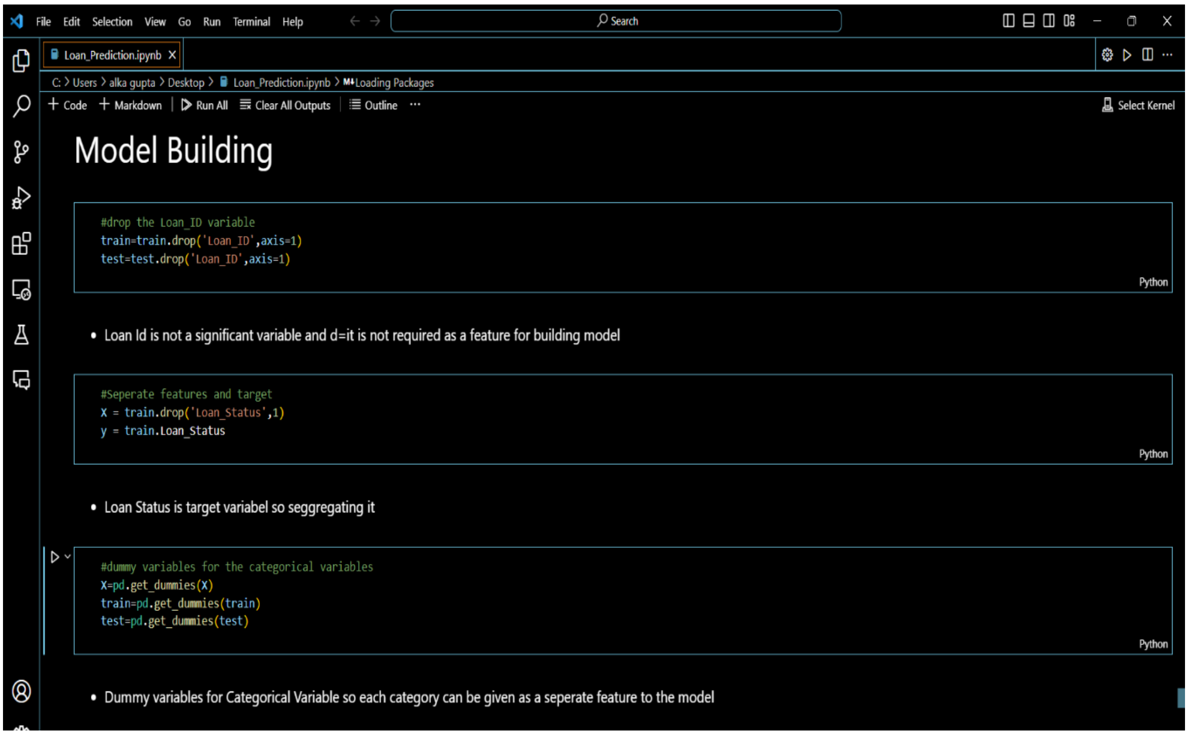
Bivariate Analysis

Categorical Variable vs Target Variable

```
Gender=pd.crosstab(train['Gender'],train['Loan_Status'])
Gender.div(Gender.sum(1).astype(float), axis=0).plot(kind="bar", stacked=True, figsize=(4,4))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a24a10410>
```

- It can be inferred that the proportion of male and female applicants is more or less **same** for both approved and unapproved loans.


File Edit Selection View Go Run Terminal Help
Search
Loan_Prediction.ipynb X
C: > Users > alka gupta > Desktop > Loan_Prediction.ipynb > M↓Loading Packages
+ Code + Markdown | Run All Clear All Outputs | Outline ···
Select Kernel
Python
Python
Python

Model Building

```python
#drop the Loan_ID variable
train=train.drop('Loan_ID',axis=1)
test=test.drop('Loan_ID',axis=1)
```

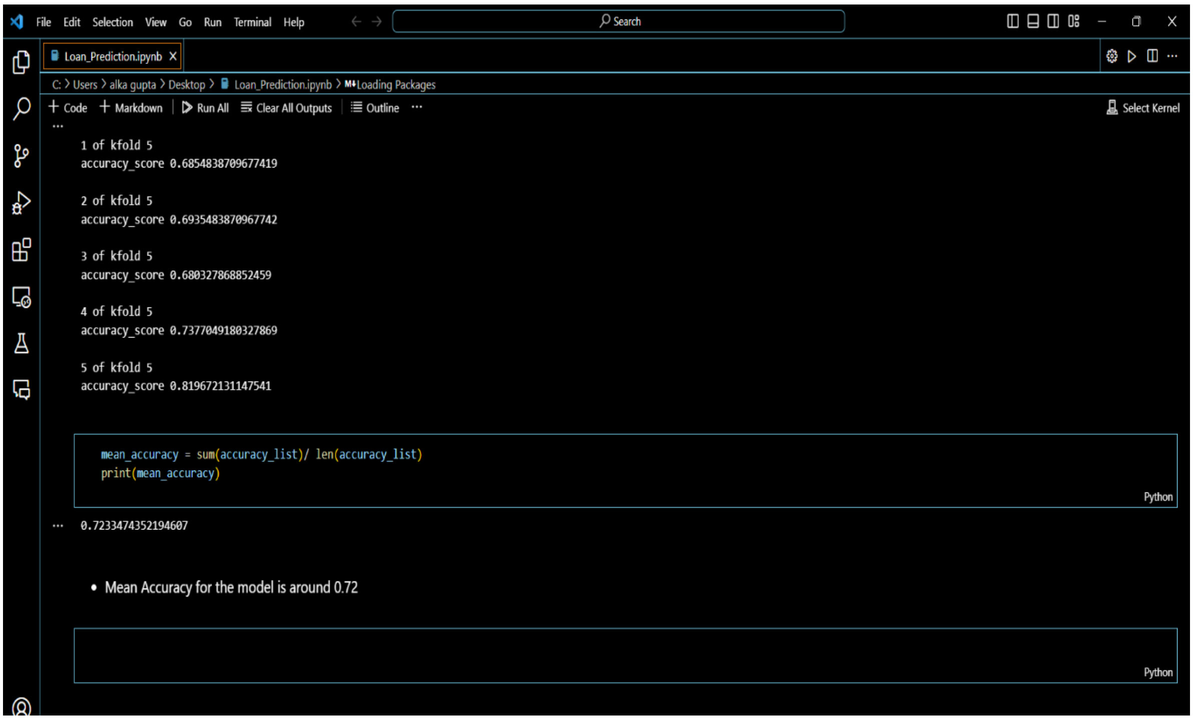
- Loan Id is not a significant variable and d=it is not required as a feature for building model

```python
#Seperate features and target
X = train.drop('Loan_Status',1)
y = train.Loan_Status
```

- Loan Status is target variabel so seggregating it

```python
#dummy variables for the categorical variables
X=pd.get_dummies(X)
train=pd.get_dummies(train)
test=pd.get_dummies(test)
```

- Dummy variables for Categorical Variable so each category can be given as a seperate feature to the model


File Edit Selection View Go Run Terminal Help
Search
Loan_Prediction.ipynb X
C: > Users > alka gupta > Desktop > Loan_Prediction.ipynb > M↓Loading Packages
+ Code + Markdown | Run All Clear All Outputs | Outline ···
Select Kernel
Python
Python

```
1 of kfold 5
accuracy_score 0.6854838709677419

2 of kfold 5
accuracy_score 0.6935483870967742

3 of kfold 5
accuracy_score 0.680327868852459

4 of kfold 5
accuracy_score 0.7377049180327869

5 of kfold 5
accuracy_score 0.819672131147541
```

```python
mean_accuracy = sum(accuracy_list)/ len(accuracy_list)
print(mean_accuracy)
```

```
0.7233474352194607
```

- Mean Accuracy for the model is around 0.72

CONCLUSION

In conclusion, machine learning algorithms, such as Logistic regression, show promising results for modern loan approval prediction. These models can automate the decision-making process, improve efficiency, and reduce human bias. However, it is important to continuously monitor and update with new data to maintain their accuracy and relevance in dynamic financial environments.

Feature importance analysis reveals that income, credit history, and employment status are the most significant factors influencing loan approval decisions. These findings align with conventional leading practices, indicating that machine-learning models can effectively learn from historical data and replicate human decision-making processes. Despite the promising results, it is important to note that the performance of machine learning models heavily relies on the quality and the representativeness of the training data.

REFERENCES

1. https://internshala.com/jobs/internships/
2. https://www.researchgate.net/publication/357449126_THE_LOAN_PREDICTION_USING_MACHINE_LEARNING
3. https://www.javatpoint.com/